Method for Free-Energy Calculations Using Iterative Techniques

SHANKAR KUMAR,* PHILIP W. PAYNE, and MAXIMILIANO VÁSQUEZ

Protein Design Labs, 2375 Garcia Avenue, Mountain View, California 94043

Received 7 August 1995; accepted 10 October 1995

ABSTRACT

We present here a new iterative technique for reliable estimation of multidimensional free energy and potential of mean force (PMF) values by computer simulation. This method is an extension of the weighted histogram analysis method [S. Kumar et al., *J. Comp. Chem.*, **13**, 1011, (1992)]. We have tested the technique by generating free-energy-based Ramachandran plots and by computing the PMF values for end-to-end distances for several polypeptides using the ECEPP/2 and AMBER force fields.

Introduction

The size of representative ensembles required for accurate estimation of free-energy or potential of mean force (PMF) values from computer simulations makes such calculations very expensive. To obtain just the qualitative features of a free-energy map is often difficult. Because such calculations are important for the elucidation of structure–function relationships in biomolecular interactions and rational drug design, a great deal of research effort has been expended on making these calculations more efficient and accurate.

We present here a robust and efficient iterative technique to reliably estimate multidimensional free-energy or PMF values by computer simulations. We will show that this method can be used to compute reliable and accurate free-energy maps that exhibit great variations (> 30 kcal/mol) from region to region. Furthermore, this can be accomplished by using computer time very efficiently. Comparison with studies in our laboratory that used conventional sampling techniques shows that the new technique achieves about an order of magnitude speed-up in the free-energy calculations, while at the same time enhancing the precision of the results.

The single histogram equations proposed by Salsburg et al.[1] in 1959 were the first method to improve the efficiency of calculating ensemble averages. Although the single histogram equations were not developed in the context of biomolecular properties,[1,2] the method is nonetheless very relevant to biomolecular simulations.[3,4] The technique we present here generates distributions iteratively

*Author to whom all correspondence should be addressed.

CCC 0192-8651 / 96 / 101269-07

in the space of any molecular property, $\tilde{\xi}$, of interest. These distributions can then be used to accurately estimate PMF or free-energy values as a function of $\tilde{\xi}$. The tilde over ξ denotes the possible multidimensional nature of the parameter. In some cases, the method presented here can be used to calculate free energies from just the single histogram equations,[2,3] i.e., effectively, from just one simulation.

The ideas behind the generation of distributions so as to obtain thermodynamic parameters have until now emphasized changing the inverse temperature β so as to get a flatter distribution in the energy E [or Hamiltonian $H(q_i)$ where the q_i denote the coordinates of the system]. Simulations in the multicanonical ensemble of Berg and coworkers[5,6] illustrate this approach. Sometimes a biasing weight is also added and the partition function is expressed as a sum over the various ensembles.[7]

In this article, we create a nearly flat distribution in the space spanned by the coordinate $\tilde{\xi}$. A flat distribution implies a uniform error distribution in the histograms. Thus, if the goal is to estimate a free-energy-based ϕ–ψ Ramachandran map[8] for a peptide, then one needs to first generate a nearly flat distribution in the ϕ–ψ space of interest, or if the quantity of interest is the end-to-end distance EED, then one needs to generate a flat distribution in EED space and so on. The method can also be used to generate sets of very low energy structures[4,9] and to study folding transitions in proteins.[10]

Method

OVERVIEW

The method generates distributions iteratively. The domain of interest of the $\tilde{\xi}$ surface is subdivided into N bins. The partition function, Z_i, derived from the ith (biased) distribution is:

$$Z_i \propto \int \exp\left(-\beta\left[H(q_i) + \Xi_i(\tilde{\xi})\right]\right)\left\{\prod_k \mathrm{d}q_k\right\} \quad (1)$$

with an appropriate choice of the bias or weight function $\Xi_i(\tilde{\xi})$. The weight function $\Xi_i(\tilde{\xi})$ for the ith iteration is defined as follows:

$$\Xi_i(\xi) = W_{i,m} \quad \text{(for } \xi \text{ belonging to the } m\text{th bin)}$$

$W_{i,m}$ is chosen to be the negative of the free energy associated with the mth bin. This free energy is estimated from all the overlapping distributions obtained from the preceding $(i-1)$ iterations. By overlapping, we mean that the $(i-1)$th iteration has at least one populated bin which is also populated in at least one of the previous distributions. In turn, populated is defined as having a statistically significant count, about 20 in our calculations. Since Ξ_i is constructed to be a function of $\tilde{\xi}$ only, $W_{i,m}$ is estimated from the $(i-1)$ previous iterations using the weighted histogram analysis method[3,11] as outlined below.

After $(i-1)$ iterations, estimates of the (unnormalized) probability $P_n''(\tilde{\xi}_m)(1 \le n \le i-1)$ are made from iterating the following equations[3]:

$$P_n''(\tilde{\xi}_m) = \frac{\sum_{k=1}^{i-1}{}'' N_k(\tilde{\xi}_m)\exp[-\beta W_{n,m}]}{\sum_{p=1}^{i-1}{}'' n_p \exp[f_p - \beta W_{p,m}]} \quad (2)$$

and

$$\exp[-f_j] = \sum_m P_j''(\tilde{\xi}_m) \quad (3)$$

Here, double primes mean that the summation and the probabilities are relevant only for distributions that are "connected" (by direct or indirect overlaps). $\tilde{\xi}_m$ denotes a value of $\tilde{\xi}$ that falls in the mth bin, n_p is the number of data points archived in the pth iteration, and $N_k(\tilde{\xi}_m)$ is the histogram of $\tilde{\xi}_m$ during the kth iteration. f_j is the free energy associated with the jth iteration. Details of iterating eqs. (2) and (3) have been described elsewhere.[3] With the free energies f_j thus determined, the values of $W_{i,m}$ are determined from the following relation:

$$\exp[\beta W_{i,m}] = \frac{\sum_{k=1}^{i-1}{}'' N_k(\tilde{\xi}_m)}{\sum_{p=1}^{i-1}{}'' n_p \exp[f_p - \beta W_{p,m}]} \quad (4)$$

As only differences in the $W_{i,m}$ are meaningful, the zero level can be arbitrarily assigned.

SPECIAL CASES

Some special circumstances cause us to modify the equations given above. First, if the histogram counts are low, i.e., below some predefined num-

ber, η, for some $\tilde{\xi}_m$, then the $W_{i,m}$ are modified to $W'_{i,m}$ as follows:

$$W'_{i,m} = \begin{cases} W_{i,m} - \Delta & \text{if } \sum_{k=1}^{i-1}{}'' N_k(\tilde{\xi}_m) < \eta \\ W_{i,m} & \text{if } \sum_{k=1}^{i-1}{}'' N_k(\tilde{\xi}_m) \geq \eta \end{cases} \quad (5)$$

Here, Δ is an arbitrarily chosen value (generally of the order of 1.0 kcal/mol) so as to enhance sampling for $\tilde{\xi} = \tilde{\xi}_m$ and can be adjusted dynamically if necessary.

Second, if the $(i-1)$th iteration is disconnected from all of the preceding simulations then $W_{i,m}$ is determined as follows:

$$W_{i,m} = \begin{cases} \frac{1}{\beta}\ln\left[\frac{N_{i-1}(\tilde{\xi}_m)}{N_{i-1}^{max}}\right] + W_{i-1,m} \\ \qquad \left(\text{if } N_{i-1}(\tilde{\xi}_m) \geq \eta\right) \\ \frac{1}{\beta}\ln\left[\frac{N_{i-1}^{min}}{N_{i-1}^{max}}\right] + W_{i-1,m} - \Delta \\ \qquad \left(\text{if } N_{i-1}(\tilde{\xi}_m) < \eta\right) \end{cases} \quad (6)$$

where N_j^{max} is the maximum value taken by the histogram in the jth iteration (say, at $\tilde{\xi}_j^{max}$), and N_j^{min} is the minimum value among the subset of histogram values that also exceed some predefined η during the jth iteration (say, at $\tilde{\xi}_j^{min}$). It should be noted that a "disconnected" iteration could eventually become "connected" due to a subsequent "bridging" distribution.

CONVERGENCE

The iteration process converges when all the bins have nonzero occupancy and $\Xi_j(\tilde{\xi}_j^{min})$ is less than some preassigned value W_{max}. A value of W_{max} between 1.0 and 2.25 kcal/mol works very well for biomolecular systems described by molecular mechanics type potentials. At room temperature, these values of W_{max} indicate that the ratio of the most populated to the least populated bin is between about 5 and 44, thus explaining what we mean by (nearly) "flat" distributions. Let us call the converged weights $\Xi(\tilde{\xi}_i)$. A final long simulation is then carried out with the converged weights and at temperature $T = 1/\beta$. The probabilities are then given by the simple expression[2]:

$$P(\tilde{\xi}_i) = \frac{N(\tilde{\xi}_i)\exp[\beta\Xi(\tilde{\xi}_i)]}{\sum_j N(\tilde{\xi}_j)\exp[\beta\Xi(\tilde{\xi}_j)]} \quad (7)$$

From the probabilities thus calculated the PMFs can be directly determined.

The convergence criteria given above are too stringent for many cases. It is not always possible to fill all the bins that partition the $\tilde{\xi}$ surface. In such a case, the iterations are considered to have converged if the absolute difference $|W_{i,m} - W_{i-1,m}|$ is less than some predefined ε ($\varepsilon > 0$) for all values of m and if $\Sigma_k N_k(\tilde{\xi}_m)$ is greater than some predefined number (thus ensuring good sampling). In this case, the negative of the free energies are the (converged) values of the weights, $W_{i,m}$.

To summarize, the method uses a feedback mechanism whereby information from simulations already carried out is used to calculate a "best" estimate of the "weights" (or the free energies that one is interested in). These weights, in turn, are used to bias the subsequent simulation. This procedure is carried out iteratively until the convergence criteria are met.

Applications

To vigorously test the procedure outlined above, we used this iterative scheme to generate PMF profiles for the end-to-end distance of deca-glycine and deca-alanine using the ECEPP/2[12-14] and AMBER[15] force fields. Simulations were conducted using Monte Carlo procedures.[16,17] We also used this procedure to generate free-energy-based Ramachandran (ϕ, ψ) plots using both ECEPP/2 and AMBER potentials for blocked alanine.

The reason for choosing to map the PMF of these parameters, i.e., the two torsion angles in a Ramachandran plot and the end-to-end distance, were twofold. First, we had to establish that this method provided a viable alternative to conventional techniques. Second, having established the first point, we wanted to subject the method to the severe test of mapping free-energy surfaces that exhibited great variations (> 10.0 kcal/mol) from region to region. Thus, we chose to work with the ECEPP/2 force field, which is a rather rigid energy surface. The PMF of the polypeptide end-to-end distance is also a challenging calculation be-

cause it involves a "complex projection" from a multidimensional to a one-dimensional surface.

The end-to-end distance is a descriptive parameter commonly used in polymer physical chemistry.[18] We carried out three independent calculations of the PMF of the end-to-end distance for $(Gly)_{10}$ and for $(Ala)_{10}$ using ECEPP/2 and AMBER. In each case, the end-to-end distance is defined as the distance between the first and tenth α-carbon atoms. The results for $(Gly)_{10}$ are shown in Figure 1. The ECEPP/2 curve is shown as an average over three independent runs, for a total of 47 iterations, with the error bars representing maximum deviations from the individual runs. Each iteration was composed of 40,000 sweeps (one sweep is defined as one Monte Carlo move for every atom [AMBER] or for every torsion angle [ECEPP/2] in the system) with the first 1000 sweeps being equilibration steps. The error bars indicate that the PMF from the individual runs show good agreement with each other. The greatest differences in the PMF values are in the very short distance range. A simple explanation for this behavior comes from the entropic effect arising from having very many different conformations, with perhaps widely different internal energies, which may map to short (end-to-end) distances; in other words, the density of states is quite large for short distances.[19] This is not the case for long distances, where very few conformations map to the extreme values. This effect may be understood qualitatively by reference to an idealized Gaussian model; for example, the radial distribution for this model decays rapidly for end-to-end distances larger than the root mean square for the model ($n^{1/2}L$ where n is the number of bonds in the chain, and L is the length of each bond); we may also refer to an improved treatment based on the more realistic freely jointed chain model.[20]

Figure 1 also shows the end-to-end distance PMF for $(Gly)_{10}$ with AMBER. The PMF graph was calculated from combining three separate runs (for a total of 131 iterations) by using eqs. (2)–(4). Each iteration comprised 200,000 sweeps with the first 2000 in each iteration being equilibration steps. The errors were calculated as the maximum deviation from the PMF graphed and the PMF estimated from the individual runs. While the errors are only approximate, they nevertheless represent a reliable order of magnitude estimate. As with the ECEPP/2 model, the largest errors (about ±1 kcal/mol) occur in the very short end-to-end distance region. The minimum of the PMF occurs at 14 Å which is characteristic of the α-helix.

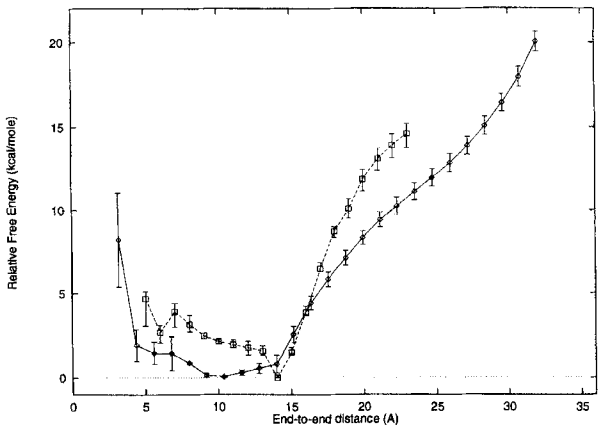

FIGURE 1. Potential of mean force of end-to-end distance of $(Gly)_{10}$. The solid line corresponds to the ECEPP/2 potential and the dashed line to AMBER.

Figure 2 shows the rate of convergence for different ranges of the $(Gly)_{10}$ end-to-end distance space for one of the calculations with ECEPP/2. It is clear that, after just a few iterations, the qualitative features of the lowest free-energy regions as well as of the long-distance regime are becoming well established.

The calculations on $(Ala)_{10}$ with ECEPP/2 (Fig. 3) yield qualitatively similar results as $(Gly)_{10}$ with ECEPP/2. The end-to-end distance (between 3.2 and 30.0 Å) were divided into 150 bins and each iteration comprised 40,000 sweeps with the first 1000 being equilibration steps. There were 25 iterations in each run. As with $(Gly)_{10}$ there is a broad low energy region spanning about 8 Å with the minimum being around 9 Å. The end-to-end distance PMF for $(Ala)_{10}$ with AMBER is shown in

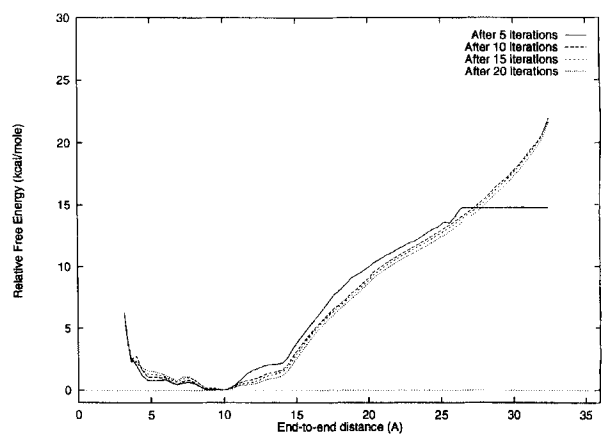

FIGURE 2. The convergence of the potential of mean force of the end-to-end distance of $(Gly)_{10}$ with the ECEPP/2 potential.

Figure 3. This is qualitatively similar to the AMBER PMF for $(Gly)_{10}$ shown in Figure 1; however, the α-helical minimum is much deeper here.

It is instructive to compare the results obtained for different potential energy models, AMBER or ECEPP/2, of the same polypeptide. The behavior of the end-to-end distance for deca-glycine is roughly similar in each model. Both the PMFs are characterized by a broad flat energy region extending for about 7 Å (between 6 and 13 Å). The global minimum obtained with ECEPP/2 (9 Å) is separated by about 5 Å from the global minimum obtained with AMBER (14 Å). However, deca-alanine behaves quite differently. Deca-alanine in AMBER has a deep free-energy minimum (with a well depth of about 6 kcal/mol) at a distance that corresponds to the α-helical conformation. While deca-alanine in ECEPP/2 is also believed to have a global minimum at the α-helix (the work of Ripoll and co-workers[21] suggests that the global energy minimum for icosaalanine is the α-helix as it is for deca-glycine[22]), our calculations suggest that this state is not as strongly preferred when free energy is considered. One possible reason for this result is the stronger electrostatic interaction present in AMBER, which may make hydrogen-bonding interactions more stabilizing than in ECEPP/2.

The calculation of the free energy (ϕ, ψ) map for the blocked alanine peptide with the ECEPP/2 potential (Fig. 4) shows that our approach gives accurate free-energy profiles even for very high energy (> 30 kcal/mol) regions. For this simple system we may compare results with a standard adiabatic map as calculated by Rotterman et al.[23] The major features are reproduced in both sets of calculations; i.e., the presence of low energy minima in regions corresponding to extended, C_7 equatorial, and right-handed α-helical structures; the occurrence of a relatively high energy minimum in the left-handed α-helical region; and the absence of stable structures in the C_7 axial region. It is also possible to compare results of Figure 4 with those obtained with identical methodology using the AMBER potential (Fig. 5). The comparison parallels that of Rotterman et al.[23] using adiabatic maps. For example, when going from ECEPP/2 to AMBER maps (either free energy or adiabatic) there is a relative destabilization of the α-helical minima, as well as noticeable stabilization of the C_7 axial region. It should be noted that in the case of Figure 5 the weights converged in just four iterations, thus enabling eq. (7) to be used for the PMF calculations—i.e., the two-dimensional (ϕ, ψ) map was calculated from just one simulation of moderate length.

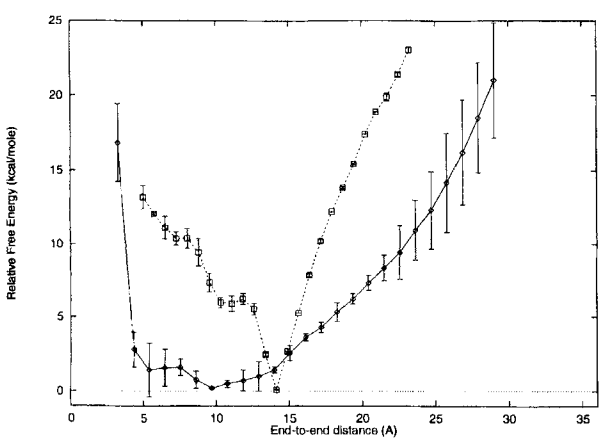

FIGURE 3. PMF of end-to-end distance of $(Ala)_{10}$. The solid line corresponds to the ECEPP / 2 potential and the dashed line to AMBER.

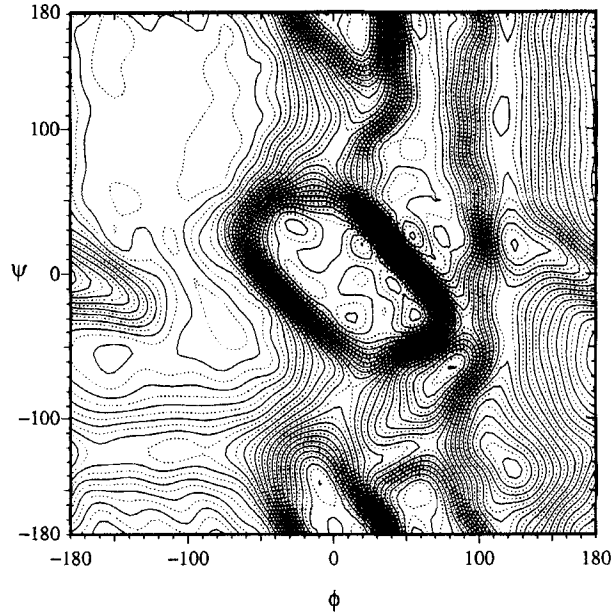

FIGURE 4. Potential of mean force of (ϕ, ψ) for blocked alanine peptide with ECEPP / 2. The lowest contour is at 0 kcal / mol and the contours are spaced 1.0 kcal / mol apart.

Conclusions

In summary, we have developed a powerful technique to compute multidimensional free-energy maps. In many cases it is possible to obtain reliable free-energy estimates from, effectively, a

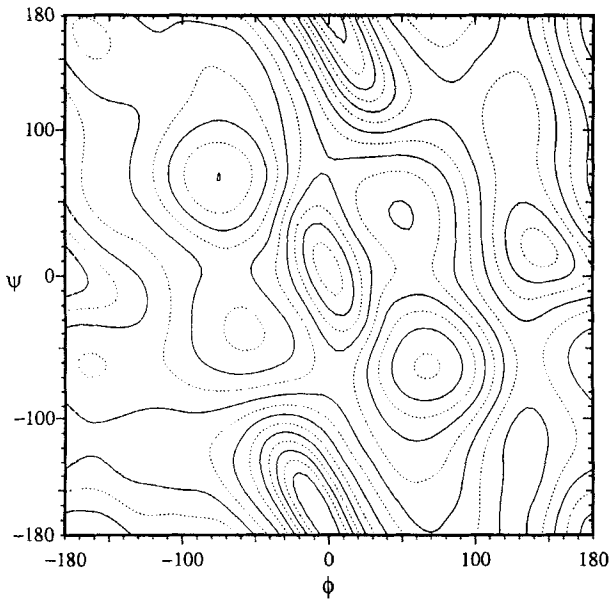

FIGURE 5. Potential of mean force of (ϕ, ψ) for blocked alanine peptide with AMBER. The lowest contour is at 0 kcal / mol and the contours are spaced 1.0 kcal / mol apart.

single simulation. The method is robust as can be seen from its ability to map large free-energy differences. Furthermore, computer time is used efficiently. For instance, to gather sufficient data to cover 400 two-dimensional bins in the calculation of a two-dimensional (ϕ, ψ) free-energy map for the central alanine in a blocked alanine tripeptide using molecular dynamics and the WHAM[3] technique took about 56 hours of CPU time. In contrast, only 14 hours were used to achieve convergence for 576 two-dimensional bins using the method described here (all calculations being carried out on a Silicon Graphics Indigo workstation). Thus this method is seen to provide greater accuracy while significantly reducing computer time.

As in the case of Berg and co-workers[5,6] the method described here employs feedback from previous simulations to improve the biasing weights of the current simulation. The optimal calculation of these weights is achieved by the WHAM[3] technique. Thus, our iterative method inherits the advantages of WHAM[3] which are:

1. By optimally linking all the overlapping iterations no data is wasted. The entire simulation is taken into account without any data having to be discarded as might happen with traditional umbrella sampling techniques.
2. The core of our iterative technique relies on the simple expression for error given by WHAM[3]; the relative error in the probability, $P(\xi)$, is inversely proportional to the square root of $\Sigma_k N_k(\xi)$. Thus, this method is able to provide reliable estimates of free energy by optimally concentrating subsequent simulations in regions of ξ where the total histogram count is low.
3. The expression for the error in the relative probabilities implies that carrying out more simulations guarantees a reduction of the relative errors.

The examples cited here used Monte Carlo techniques for performing the simulations. However, it is possible to adapt our algorithm to the problem of estimating differential binding energies using molecular dynamics techniques where until now only the classic thermodynamic perturbation methods have been used. We are currently working on these aspects and our progress will be the subject for future communications.

Acknowledgment

We are grateful to Dr. Cary Queen for reading the manuscript and for making helpful suggestions.

References

1. Z. W. Salsburg, J. D. Jacobson, W. Fickett, and W. W. Wood, *J. Chem. Phys.*, **30**, 65–72 (1959).
2. A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.*, **61**, 2635–2638 (1988).
3. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comput. Chem.*, **13**, 1011–1021 (1992).
4. U. H. E. Hansmann and Y. Okamoto, *J. Comput. Chem.*, **14**, 1333–1338 (1993).
5. B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.*, **68**, 9–12 (1992).
6. B. A. Berg, *Int. J. Modern Phys., C* **3**, 1083–1098 (1992).
7. A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, *J. Chem. Phys.*, **96**, 1776–1783 (1992).
8. G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure*, Springer-Verlag, New York, 1979.
9. U. H. E. Hansmann and Y. Okamoto, *Physica A*, **212**, 415 (1994).
10. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.*, **98**, 4940–4948 (1994).
11. S. Kumar, D. Bouzida, P. A. Kollman, R. H. Swendsen, and J. M. Rosenberg, *J. Comput. Chem.*, **16**, 1339 (1995).

12. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79,** 2361–2381 (1975).
13. G. Nemethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87,** 1883–1887 (1983).
14. M. J. Sippl, G. Nemethy, and H. A. Scheraga, *J. Phys. Chem.*, **88,** 6231–6233 (1984).
15. S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, *J. Comput. Chem.*, **7,** 230–252 (1986).
16. D. Bouzida, S. Kumar, and R. H. Swendsen, *Phys. Rev. A*, **45,** 8894–8901 (1992).
17. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.*, **21,** 1087–1092 (1953).
18. C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry*, W. H. Freeman and Co., San Francisco, CA, 1980.
19. T. F. Havel, *Biopolymers*, **29,** 1565–1585 (1990).
20. J. Kostrowicki and H. A. Scheraga, *Comp. Polym. Sci.*, **5,** 47–55 (1995).
21. D. R. Ripoll and H. A. Scheraga, *Biopolymers*, **27,** 1283–1303 (1988).
22. D. R. Ripoll, M. J. Vásquez, and H. A. Scheraga, *Biopolymers*, **31,** 319–330 (1991).
23. I. K. Rotterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, *J. Biomolec. Struct. Dyn.*, **7,** 421–453 (1989).